Maximum-Likelihood Phylogenetic Inference with Selection on Protein Folding Stability

Miguel Arenas,[1] Agustin Sánchez-Cobos,[1] and Ugo Bastolla*,[1]

[1]Department of Cell Biology and Immunology, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Universidad Autónoma de Madrid, Madrid, Spain

***Corresponding author:** E-mail: ubastolla@cbm.csic.es.

Associate editor: Jeffrey Thorne

Abstract

Despite intense work, incorporating constraints on protein native structures into the mathematical models of molecular evolution remains difficult, because most models and programs assume that protein sites evolve independently, whereas protein stability is maintained by interactions between sites. Here, we address this problem by developing a new mean-field substitution model that generates independent site-specific amino acid distributions with constraints on the stability of the native state against both unfolding and misfolding. The model depends on a background distribution of amino acids and one selection parameter that we fix maximizing the likelihood of the observed protein sequence. The analytic solution of the model shows that the main determinant of the site-specific distributions is the number of native contacts of the site and that the most variable sites are those with an intermediate number of native contacts. The mean-field models obtained, taking into account misfolded conformations, yield larger likelihood than models that only consider the native state, because their average hydrophobicity is more realistic, and they produce on the average stable sequences for most proteins. We evaluated the mean-field model with respect to empirical substitution models on 12 test data sets of different protein families. In all cases, the observed site-specific sequence profiles presented smaller Kullback–Leibler divergence from the mean-field distributions than from the empirical substitution model. Next, we obtained substitution rates combining the mean-field frequencies with an empirical substitution model. The resulting mean-field substitution model assigns larger likelihood than the empirical model to all studied families when we consider sequences with identity larger than 0.35, plausibly a condition that enforces conservation of the native structure across the family. We found that the mean-field model performs better than other structurally constrained models with similar or higher complexity. With respect to the much more complex model recently developed by Bordner and Mittelmann, which takes into account pairwise terms in the amino acid distributions and also optimizes the exchangeability matrix, our model performed worse for data with small sequence divergence but better for data with larger sequence divergence. The mean-field model has been implemented into the computer program Prot_Evol that is freely available at http://ub.cbm.uam.es/software/Prot_Evol.php.

Key words: **structurally constrained substitution models, folding stability, misfolded state, maximum-likelihood estimate.**

Introduction

A variety of amino acid substitution models of evolution have been developed to perform phylogenetic analysis. The simplest models are based on the assumption that protein sites evolve independently and identically according to empirical substitution matrices such as JTT (Jones et al. 1992) or WAG (Whelan and Goldman 2001). Despite the great success of these simple models (Yang et al. 1998), in particular for evaluating phylogenetic trees and inferring evolutionary and population parameters with the maximum-likelihood (ML) method (Whelan et al. 2001; Felsenstein 2004), they present the important drawback that they ignore the information contained in protein structures (Robinson et al. 2003; Wilke 2012).

Selection on protein folding stability ultimately acts on interactions between sites, implying that sites do not evolve independently. Nevertheless, giving up the independence among sites generates huge complications in the computation of the likelihood. Because of this reason, several groups have tried to incorporate the effect of protein structure through site-specific substitution matrices. On one extreme, Koshi and Goldstein (1998, 2001) and Koshi et al. (1999) have developed substitution models that consider physicochemical properties of amino acids. On the other extreme, Halpern and Bruno (1998) proposed to adopt different amino acid frequencies for each position of a protein; despite improving over simpler models, this approach requires very large amount of data to fit all the needed parameters. Lartillot and Philippe (2004) interpolated between these two approaches letting the number of site classes to be a parameter of their model. Instead of fixing the number of classes and their parameters, which can result in overfitting, they integrate over all of these parameters through Monte Carlo sampling. This method is often used in simulations, but it is less established than nonmixture models for phylogenetic inference because of its computational burden.

On the other hand, progresses in statistical mechanical models of protein folding (Plotkin and Onuchic 2002; Shakhnovich 2006; Chan et al. 2011) have prompted since long time models of protein evolution that enforce selection

Article

on the stability of the native state (Gutin et al. 1995; Babajide et al. 1997; Bussemaker et al. 1997; Govindarajan and Goldstein 1997; Mirny et al. 1998; Tiana et al. 1998; Bastolla et al. 1999, 2003; Bornberg-Bauer and Chan 1999; Dokholyan and Shakhnovich 2001; Parisi and Echave 2001; Taverna and Goldstein 2002; Bloom et al. 2005; DePristo et al. 2005; Goldstein 2011; Grahnen et al. 2011; Huang et al. 2014), recently reviewed in Liberles et al. (2012). Although these models are not applicable to the important class of natively unfolded proteins (Uversky and Dunker 2010), it is clear that the stability of the native state is a very important determinant of protein evolution. Simple models of protein folding allow simulating protein evolution (see Arenas et al. 2013, for a recent implementation in the context of phylogenetic trees) and they produced important insights. Nevertheless, it has been difficult to apply them for phylogenetic inference. A pioneering contribution was made by Fornasari et al. (2002), who adopted simulations of their structurally constrained protein evolution model for computing site-specific substitution matrices still assuming independent sites. A few groups abandoned the independent sites approximation, proposing substitution models that take into account pairwise amino acid distributions, in particular Rodrigue et al. (2005) and, quite recently, Bordner and Mittelmann (2014). However, the computational implementation of pairwise-sites models is complicated and it cannot be combined with standard programs for phylogenetic inference.

New Approaches: The Mean-Field Model

Here, we build on previous work by one of us and coworkers, who noted that contact-based models of protein folding combined with the assumption of independent sites and other approximations allow to analytically compute site-specific amino acid frequencies. We call this a mean-field (MF) model, because each site evolves independently but taking into account in a self-consistent way the MF generated by the other sites. Approximating contact interaction energies with their hydrophobic component, the previous model established an explicit relationship between the average hydrophobicity of a site in a family of protein sequences and its connectivity at the structural level (Bastolla et al. 2005, 2008; Porto et al. 2005) and it was later extended to generate a substitution model (Bastolla et al. 2006). The MF model that we present here builds on that proposal, but is not explicitly based on hydrophobicity and it adopts an improved representation of the statistical mechanical model of the misfolded state (Minning et al. 2013). The model generates the site-specific amino acid distributions that are maximally close to a background amino acid distribution and fulfill a constraint on the average folding free energy that effectively maintains the stability of the native state. The only free parameters of the model are the Lagrange multiplier that imposes the selective constraint and the parameters that define the background distribution. They are determined imposing that the observed protein sequence has ML with respect to the site-specific amino acid distributions. For most proteins, the resulting amino acid distributions produce sequences in which the native state is on the average stable, as assessed through our folding model, despite this condition is not explicitly imposed. Importantly, we found that considering the misfolded state produces higher likelihood, larger stability, and more realistic hydrophobicity values than only considering the native and the unfolded state. In all cases, the amino acid distributions observed in natural protein families agreed better with the MF model than with the frequencies of empirical substitution models.

We then generated site-specific matrices of substitution rates by combining the MF site-specific stationary distributions with an exchangeability matrix obtained from an empirical substitution model. We applied the resulting rate matrices for phylogenetic inference, comparing their Akaike Information Criterion (AIC) scores (Akaike 1974) (likelihood penalized by the number of free parameters) with those of other structurally constrained models of protein evolution, finding that our model produces better results than a recent model with independent sites that takes into account solvent accessibility but disregards the misfolded state (Bordner and Mittelmann 2014), and even better than a pairwise model developed by Rodrigue et al (2005) as reported in Bordner and Mittelmann (2014), despite this model explicitly considers correlations between sites. With respect to the much more complex model recently developed by Bordner and Mittelmann (BM) (2014), that takes into account pairwise terms in the amino acid distributions and optimizes more parameters than our model, we obtained worse results for three protein families with very small sequence divergence but better results for one family with larger sequence divergence.

Finally, we examined eight highly divergent protein families obtained from the Pfam database (Punta et al. 2012). For most of them, the likelihood of our model was higher than one of the empirical model. For those cases in which the empirical model gave better performances, the MF model became superior if we eliminated proteins with sequence identity smaller than 35% with respect to the representative structure, consistent with the fact that proteins with low sequence identity may have divergent structures.

Overall, the MF model provides a structure-based modeling of protein evolution that considers the misfolded state, and it allows a fast computation of the evolutionary parameters per site that can be easily applied to phylogenetic inference. The MF model associates to a known protein structure a probability distribution in sequence space with the following properties:

1) The global probability distribution of a protein family is modeled as the product of amino acid distributions of single sites, that is, sites are considered independent:

$$P(a_1 \cdots a_L) = \prod_{i=1}^{L} P^{\mathrm{MF},i}_{a_i}, \qquad (1)$$

where i labels any of the L sites and a_i labels the amino acid at site i. The assumption of site-independent

evolution is necessary for computationally efficient algorithms, such as the most commonly used methods for phylogenetic inference (Felsenstein 2004).

2) The single-site amino acid distributions are the product of the site-independent background distribution determined by the mutation process P_a^{mut} times site-specific selection factors $f_a^{\mathrm{sel},i}$,

$$P_a^{\mathrm{MF},i} = P_a^{\mathrm{mut}} f_a^{\mathrm{sel},i}. \quad (2)$$

3) The background distribution P_a^{mut} is modeled in two different ways: Either 1) the amino acid frequencies are treated as $m = 19$ free parameters (the 20th parameter is determined through the normalization condition) or 2) the amino acid frequencies are obtained from a codon-based substitution model that has $m = 4$ free parameters (three nucleotide frequencies and the transition–transversion ratio). This model is selectively neutral, except that stop codons are forbidden, and the amino acid frequencies are obtained as the sum of the stationary frequencies of their codons (see also Methods). In both cases, the free parameters are determined by maximizing the likelihood of the amino acid frequencies observed in the protein structure plus those present in a protein family, if they are available. In case (1), this simply means that we equate the background frequencies and the observed frequencies.

4) The selection factors $f_a^{\mathrm{sel},i}$ are determined imposing that the resulting global distribution presents minimum Kullback–Leibler (KL) divergence with respect to the background distribution P_a^{mut}, for given average folding free energy ΔG. If we impose this constraint through a Lagrange multiplier Λ, the $20L$ parameters $f_a^{\mathrm{sel},i}$ are determined by minimizing the quantity

$$\sum_i \sum_a P_a^{\mathrm{MF},i}[\log(P_a^{\mathrm{MF},i}) - \log(P_a^{\mathrm{mut}}) + z_i] + \Lambda \sum_{a_1 \cdots a_L} P_{a_1}^{\mathrm{MF},1} \cdots P_{a_L}^{\mathrm{MF},L} \Delta G(C_{\mathrm{nat}}, a_1 \cdots a_L). \quad (3)$$

where z_i is the Lagrange multiplier that imposes the normalization constraint $\sum_a P_a^{\mathrm{MF},i} = 1$, C_{nat} is the contact matrix of the native structure, and ΔG is the folding free energy of the native state in the sequence $a_1 \cdots a_L$. The sum is over all possible sequences of L amino acids. Although this is an astronomic number, the sum can be analytically computed exploiting the independence of each site. Note that the $f_a^{\mathrm{sel},i}$ are not free parameters, because they are completely determined by the native structure, by the properties of the misfolded ensemble, and by Λ, the multiplier that imposes the constraint on the average folding free energy. Λ is treated as a free parameter that is fixed through the condition that the model maximizes the likelihood of the protein sequence in the Protein Data Bank (PDB), $A_1 \cdots A_L$:

$$\Lambda = \mathrm{argmax}\left(\sum_i \log\left(P_{A_i}^{\mathrm{MF},\Lambda,i}\right) \right). \quad (4)$$

The minimum KL condition (eq. 3) is analogous to the condition that determines the Boltzmann distribution in statistical mechanics as the maximum entropy distribution with given average energy. Berg et al. (2004) and Sella and Hirsch (2005) showed that several models of evolutionary genetics are formally equivalent to statistical mechanics in the space of biological sequences, with minus fitness playing the role of energy and the inverse of population size playing the role of temperature. The minimum KL condition is equivalent to the maximum entropy condition if the background distribution due to mutation assigns equal probability to all amino acids, and it generalizes it for more realistic background distributions. In qualitative terms, minimum KL with respect to the mutational distribution means that selection produces the minimum possible deviation from what would be achieved by mutation alone, that is, that the selective pressure is minimal.

The evolutionary model requires to constrain the fitness, which is often modeled as the probability that the protein is in the native state, that is, $F = \mathrm{e}^{-\Delta G/kT}/\left(1 + e^{-\Delta G/kT}\right)$ (Goldstein 2011). Constraining the fitness represents the important saturation effect that evolution becomes more tolerant to deleterious mutations and effectively neutral if ΔG is very negative (Taverna and Goldstein 2002). However, the iterative procedure that we developed for computing the MF distribution has convergence problems if we constrain the fitness F exactly because of this reason: For large proteins, the fitness becomes almost a binary variable with values zero or one, and the iterative algorithm cycles between these two states. To avoid this problem, in equation (3) we resort to the better behaved approximation to constrain ΔG. Note that constraining fitness and constraining ΔG would be equivalent if the derivative of the fitness with respect to ΔG could be treated as a nonfluctuating variable.

Instead of determining the selection parameter Λ by imposing an experimentally determined value of the average folding free energy $\overline{\Delta G}$ (the average is taken over the MF distribution), we determine it with the ML condition and from this we obtain the complete amino acid distribution and compute $\overline{\Delta G}$. It is remarkable that for most proteins the obtained $\overline{\Delta G}$ is negative, that is, sequences described by the MF distribution are on the average stable, and its value is similar to the value $\Delta G(A_1 \cdots A_L)$ computed for the native protein in the PDB. This is not trivial, because smaller values of Λ produce MF models with $\overline{\Delta G} > 0$ and larger values of Λ produce MF models with too negative $\overline{\Delta G}$.

In practice, it is very cumbersome to maximize the likelihood with respect to all parameters, and we resort to approximations that allow computing the MF distribution in a time that ranges from seconds to few minutes depending on the target protein. The steps of the algorithm are described in detail in the Methods section and in the Appendix.

To apply the site-dependent amino acid distributions to phylogenetic inference, we have to construct a substitution rate matrix that has these distributions as limit distributions. As most programs for phylogenetic inference do, we assume detailed balance and choose a symmetric exchangeability

matrix E^{MF}_{ab} that determines the site-specific substitution rates $Q^{MF,i}$ as

$$Q^{MF,i}_{ab} = E^{MF}_{ab} P^{MF,i}_{b} \; a \neq b. \quad (5)$$

This ansatz automatically satisfies the detailed balance $P^{MF,i}_{a} Q^{MF,i}_{ab} = P^{MF,i}_{b} Q^{MF,i}_{ba}$, which implies that $P^{MF,i}_{a}$ is the limit distribution. The diagonal elements are determined through the normalization condition $Q^{MF,(i)}_{aa} = -\sum_b Q^{MF,i}_{ab}$.

The symmetric exchangeability matrix has 190 free parameters. We did not attempt to determine an exchangeability matrix optimally suitable for our MF model, as it was made for instance in Bordner and Mittelmann (2014). This optimization may give room to large improvement of the results, because we observed that E strongly influences the resulting likelihood. Instead, the results presented in this work are based on an exchangeability matrix derived from an empirical substitution model such as WAG or JTT, with parameters E^{emp}_{ab} and f^{emp}_{a}, according to one of the three following possible schemes:

1) The simplest choice (here denoted E) is to impose that the exchangeability matrix is the same as the empirical model, $E^{MF,E}_{ab} = E^{emp}_{ab}$. However, we expect that this choice is not optimal because empirical substitution matrices represent both mutation and selection, whereas we need an exchangeability matrix that represents only mutation, because selection is modeled through the condition on $\overline{\Delta G}$.
2) The second option, denoted as F, imposes that the site-averaged amino acid flux is the same as for the empirical model,

$$E^{MF,F}_{ab} \sum_i P^{MF,i}_{a} P^{MF,i}_{b} / L = E^{emp}_{ab} f^{emp}_{a} f^{emp}_{b}. \quad (6)$$

3) The third possibility, here denoted as Q, requires that the rate matrices $Q = Ef$ of the MF model and the empirical model are as similar as possible. Because the rate matrix uniquely determines the stationary frequencies, it is not possible that the two matrices are equal, and we impose that they are most similar in the mean-square sense. This condition requires that the symmetric parts of the rate matrices are equal:

$$E^{MF,Q}_{ab} \sum_i \left(P^{MF,i}_{a} + P^{MF,i}_{b}\right)/L = E^{emp}_{ab}\left(f^{emp}_{a} + f^{emp}_{b}\right). \quad (7)$$

The MF model has been implemented into the computer program Prot_Evol that is freely available at http://ub.cbm.uam.es/software/Prot_Evol.php (last accessed April 13, 2015). This program can analyze any protein structure in the PDB with or without a protein sequence data set in a few seconds/minutes, producing as output the site-specific amino acid frequencies and the exchangeability matrix that define the substitution process together with information on the likelihood of the native sequence with respect to the model and the computed mean folding free energy of the MF model and the native protein.

We compute the likelihood of the substitution model in two steps. In the first step, we use a global average substitution matrix $Q^{MF,\Omega}_{ab} = E^{MF}_{ab} \sum_i P^{MF,i}_{b} / L$ and we obtain optimal branch lengths for all sites with the PAML program (Yang 2007), conveniently modified (see Methods). In the second step, we run PAML for each site separately with the fixed branch lengths obtained in the first step. Note that this procedure only approximately achieves branch lengths that optimize the sum of the log likelihood of all sites.

Results

Assessment of the Mean-Field Model with Individual Proteins

In this section, we apply the MF model to a test set of 380 monomeric globular proteins in the PDB whose structure was determined through X-ray crystallography. The background distribution was obtained from the amino acid frequencies in the PDB sequence.

Site Specificity Is Determined by the Number of Contacts

We found that the properties of the site-specific distributions strongly depend on the number of native contacts of each site. As predicted (see eq. 14 and Porto et al. 2005), the average hydrophobicity $\overline{h}_i = \sum_a P^{MF,i}_{a} h(a)$ of the MF distributions is strongly negatively correlated with the number of contacts (fig. 1*A*). This is not surprising, because buried sites with more contacts tend to be more hydrophobic. However, this is not an assumption but a result of the model, and the strength of the correlation is remarkable (the correlation coefficient is $r = 0.906$ on the average).

Figure 1*B* shows that the entropy of the distributions has a maximum for an intermediate number of native contacts, consistent with our previous prediction (Porto et al. 2005). This property of our model contrasts with the common wisdom that more exposed sites are less conserved. However, it is compatible with the observation that buried sites evolve more slowly than exposed sites (Franzosa and Xia 2009). We indeed reproduced this observation using exchangeability matrices derived both from an empirical substitution process and from a mutation process (data not shown). The apparent contradiction is explained by the fact that in our model exposed sites are less variable than sites of intermediate exposure, but for proper choices of the exchangeability matrix they are characterized by a higher exchangeability rate although the number of allowed amino acids is smaller. We will discuss in detail this important aspect in a forthcoming publication.

Considering Misfolding Improves the Likelihood and Yields Stable Proteins with Realistic Hydrophobicity

We plot in figure 2*A* the log likelihood per site of the PDB sequence with respect to different MF models that we denote here by $P^{MFk,i}$, where k labels the MF model. Each point represents a protein. The likelihood of the mutation model P^{mut}_{a}, which is equal to minus the entropy of the PDB sequence, is used as a reference on the x-axis. The second type of model, $k = 0$, computes ΔG considering only the native and the unfolded state. The model with $k = 1$ considers the first moment of the contact energy of the misfolded state, $\sum_{i<j} \langle C_{ij} \rangle U(A_i, A_j)$. The model $k = 2$ also includes the second moment of the energy of the misfolded state, that

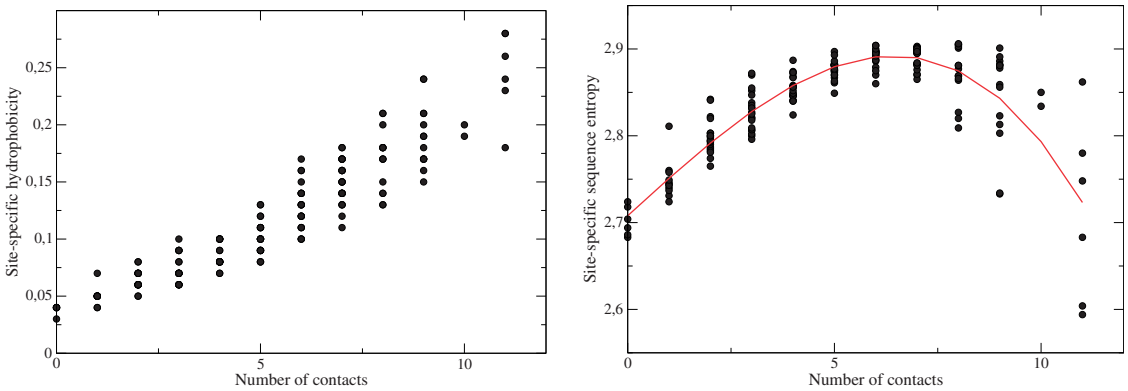

FIG. 1. Site-specific average hydrophobicity (left) and entropy (right) of the MF distributions as a function of the number of native contacts for the protein with PDB code 153l. As expected, there is a very strong correlation between hydrophobicity and number of contacts and the entropy reaches a maximum at an intermediate number of contacts.

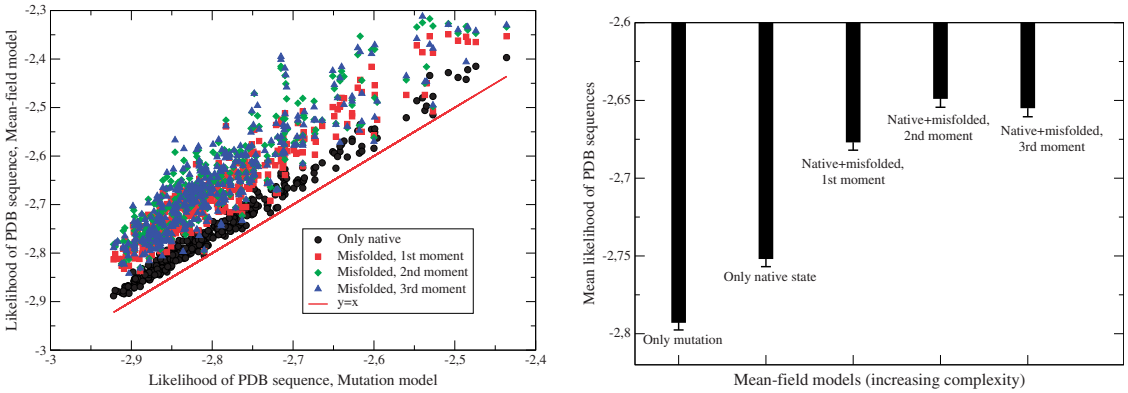

FIG. 2. Left: Log likelihood of various MF models as a function of the log likelihood of the purely mutation model. Each point represents a protein. Right: Mean log likelihood of the five types of MF models. The plotted statistical errors show that differences are significant except for the two rightmost bars.

is, the full equation (10), and the last one, $k = 3$, also includes the third moment of the misfolded energy.

We compare these different MF models in figure 2*B*, which shows the log likelihood per site averaged over all proteins for five different models. The number of parameters of the four models with selection is the same, just one more than for the purely mutational model, which corresponds to $\Lambda = 0$. Such an extra parameter yields a negligible correction to the AIC per site. One can see that the mean likelihood clearly improves going from the mutation model to the model that only takes into account the native state, and an even larger improvement is obtained considering the misfolding ensemble, which is not considered by other structurally constrained models for phylogenetic inference due to its computational complexity. The best results are obtained considering the second moment of the energy of the misfolded ensemble, while the third moment slightly worsens the results, probably due to the crude approximations needed to efficiently compute it. Therefore, in the following we adopt the model with the second moment.

In figure 3*A*, we see that the models $P_a^{\mathrm{MF2},i}$ and $P_a^{\mathrm{MF3},i}$ have more realistic average hydrophobicities for all proteins, which contribute to their higher likelihood, whereas the models $P_a^{\mathrm{MF0},i}$ (only native) and $P_a^{\mathrm{MF1},i}$ tend to have hydrophobicity larger than that of the sequence in the PDB. As a result, the average folding free energy $\overline{\Delta G}$ is positive for the model based only on the native state, in which the misfolded ensemble has lower free energy than the native ensemble, and for the mutation model that lacks site specificity, that is, the protein families described by these models are on the average unstable (fig. 3*B*). On the contrary, the models $P_a^{\mathrm{MF2},i}$ and $P_a^{\mathrm{MF3},i}$ (not shown) yield folding stability to most protein families. This is remarkable because the selection parameter Λ is fixed through the ML criterion, which does not require $\overline{\Delta G} < 0$. We found that, for the proteins for which $\overline{\Delta G} > 0$ with the model $P_a^{\mathrm{MF2},i}$, a value of Λ slightly larger than the ML produces a stable protein family. The same is not always true with the native-only model $P_a^{\mathrm{MF0},i}$.

Finally, our results depend on the temperature at which the thermodynamic computations are performed. This temperature has arbitrary units, set by the units of the contact free energy function that we adopt. We can use the mean likelihood of the proteins in the test set with respect to the model $P_a^{\mathrm{MF2},i}$ to determine the temperature parameter that yields optimal results, and that we interpret as the room temperature expressed in units of contact interactions. This optimal temperature turns out to be $T = 0.5$ (fig. 4). For this value of the temperature, the model optimally describes

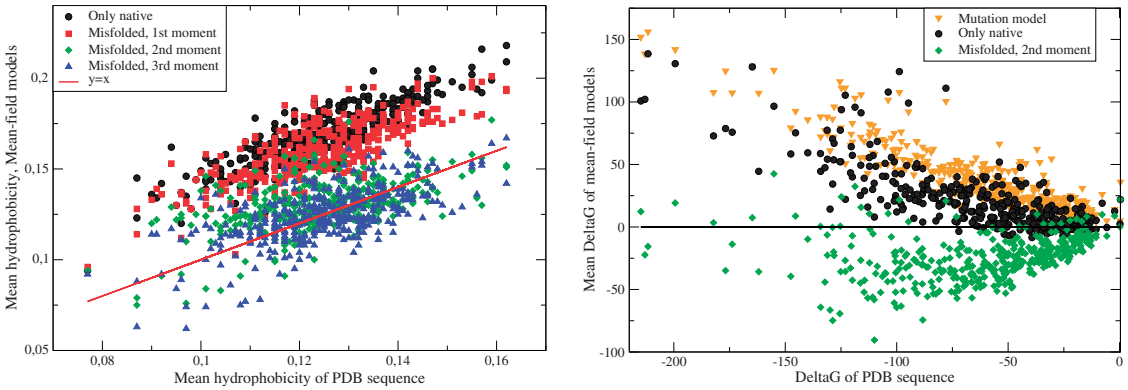

Fig. 3. Left: Average hydrophobicity of the MF models versus the average hydrophobicity of the PDB sequence. Each point represents a protein. Right: Average folding free energy (native minus misfolded) $\overline{\Delta G}$ of the MF models versus the average folding free energy of the PDB sequence. $\overline{\Delta G} < 0$ means that the MF model describes on the average stable proteins. Each point represents a protein.

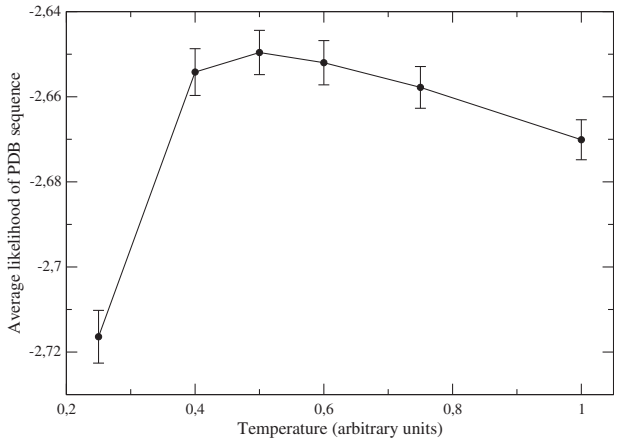

Fig. 4. Mean log-likelihood of the proteins in the test set with respect to the model $P_a^{MF2,i}$ versus the temperature in arbitrary units set by our contact interaction energy function.

protein sequences in the PDB. All reported computations are performed at this temperature.

Assessment of the Mean-Field Model with Protein Families

In this section, we compare the performances of the MF model with those of other substitution models by applying them to 12 different protein families (table 1). Four families had been previously studied by Bordner and Mittelmann (2014), so that we could directly compare the MF model with structurally constrained models presented therein, in particular the pairwise-sites substitution model based on factor graphs (hereafter BM), the independent sites model based on surface accessibility (hereafter, SA), and the pairwise model based on contact potentials that was developed by Rodrigue et al. (2005) (hereafter, RO). The remaining eight families were much more divergent than the four above (table 1), and they were randomly chosen from the seed alignments of the Pfam database (Punta et al. 2012) that possess at least ten sequences and one representative structure must be present in the PDB. In order to facilitate the comparison with the results of the RO model, we adopted an exchangeability matrix derived from the JTT model. We present results obtained with the condition F (eq. 6). Results obtained with the condition Q (eq. 7) are similar. We arbitrarily applied the WAG matrix for the other eight protein families (see below). We applied the default thermodynamic settings (i.e., temperature $T = 0.5$ and configurational entropy per residue $S_C = 0.065$) described in Methods.

Amino Acid Distributions

In order to compare the amino acid distributions observed at each site of the multiple sequence alignment with the site-dependent distributions generated by the MF model on one hand, and with the site-independent distribution adopted by the empirical model on the other hand, we measured the KL divergence at each site i:

$$d_i^{\mathrm{KL}} = \sum_a P_a^{\mathrm{obs},i}\left(\left(\log(P_a^{\mathrm{obs},i}\right) - \left(\log(P_a^{\mathrm{mod},i}\right)\right), \qquad (8)$$

where a is any of the 20 amino acids. We compute the weighted sum $D^{\mathrm{KL}} = \sum_i w_i d_i^{\mathrm{KL}}$, with weights w_i proportional to the number of aligned residues (excluding gaps) in column i of the alignment. The smaller the D^{KL}, the closer the observed and model-provided distributions are. We found that the MF model presented lower D^{KL} than the empirical model for all protein families (fig. 5), which indicates that it better represents the amino acid distributions present in the real data. Furthermore, the difference between the MF model and the empirical model increases when sequences with less than 25% sequence identity with respect to the representative protein are eliminated from the test set.

Comparison with Other Substitution Models for Phylogenetic Inference

First, we examined the four protein families studied in the recent publication by Bordner and Mittelmann (2014) (table 1). We fitted the models and computed their ML with PAML, correcting for the number of degrees of freedom (dofs) with the AIC scores, both for the MF model (19 dofs) and for the reference model JTT +G, where +G indicates heterogeneous substitution rate across sites according to a gamma

Table 1. Protein Families Collected from the Pfam Database.

Protein family	Pfam	Proteins	Uniprot	PDB	Length	⟨seq.id.⟩
Glucokinase	PF02685	4	GLK_ECO57	1SZ2	465	0.93
Homogentisate 1,2-dioxygenase	PF04209	4	HGD_HUMAN	1EY2	319	0.92
Cytochrome P450	PF00067	4	CP2A6_HUMAN	1Z10	419	0.96
Pancreatic ribonuclease	PF00074	4	RNAS1_BOVIN	1SRN	113	0.74
Triosephosphate isomerase	PF00121	56	TPIS_TRYBB	1TTI	236	0.37
Rubredoxis	PF00301	43	RUBR2_PSEOL	1R0F	53	0.45
Kinesin	PF00225	87	KAR3_YEAST	3KAR	323	0.35
Ferredoxin	PF05996	62	PCYA_SYNY3	3NB8	242	0.33
DNA ligase	PF13298	136	B1L4V6_KORCO	3P4H	118	0.46
Heat shock protein	PF00012	33	DNAK_ECOLI	2KHO	600	0.53
Oxysterol-binding protein	PF01237	153	KES1_YEAST	1ZHT	436	0.25
Retroviral aspartil protease	PF00077	50	POL_FIVPE	3OGQ	112	0.25

NOTE.—For each family, the table indicates the Pfam code, sample size, UniProt entry for a protein sequence with a PDB structure, the PDB code, number of amino acids and average sequence identity with respect to the representative protein. Note that the first four entries were selected following the study by Bordner and Mittelmann (2013) and they present a very high sequence identity.

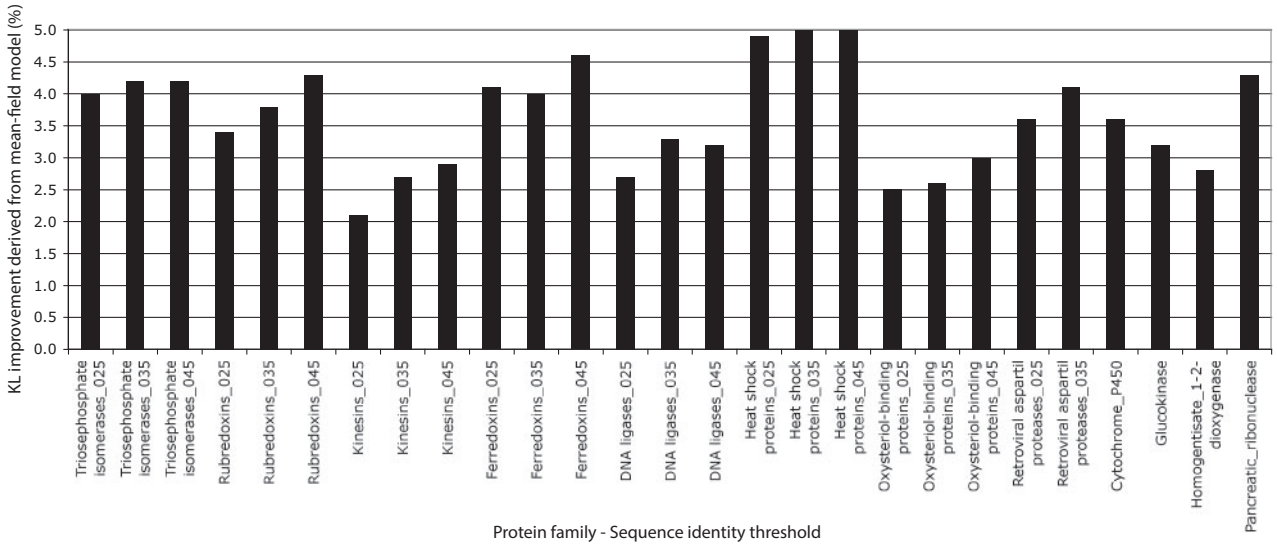

FIG. 5. Difference of KL divergence from the observed amino acid profile between the empirical model and the MF model (KLDobs_emp−KLDobs_mf) for the 12 studied protein families, under different conditions on the minimum sequence identity allowed. Positive differences mean that the observed profile agrees better with the MF model than with the empirical model.

distribution (Yang 1993) (1 dof). The results derived from the models MF, BM, SA, and RO are reported in table 2. For all of the protein families, the MF model showed a better fitting than the RO model and it was also better than the SA model for three of the four protein families, whereas the BM model presented a better fitting than the MF model for three of the four protein families. Interestingly, the MF was the best model for the family with largest divergence (average sequence identity 0.74), whereas the BM model was the best model for the other three families that present an average sequence identity larger than 0.90.

For the other eight protein families (entries 5–12 in table 1), we computed the AIC scores between MF and the WAG empirical substitution model, with 19 and 0 degrees of freedom, respectively. Here, we found a strong impact of data sets with low sequence identity on the fitting of the MF model (see sequence identities for these data sets in table 1). We explored this impact by filtering the data sets according to the sequence identity of all proteins with respect to the protein of the representative structure (too distant protein sequences are trimmed from the data set), in particular we analyzed these data sets adopting sequence identity thresholds of 0.25, 0.35, and 0.45. The AIC scores for all these data sets are presented in table 3. The results indicate that data sets with sequence identity with respect to the protein structure below 0.25 can be problematic for the MF model suggesting that distant protein sequences are poorly represented by the reference structure. All data sets with sequence identity levels

Table 2. Difference of AIC for the Structurally Constrained Substitution Models MF, BM, RO, and SA Relative to the Empirical Substitution Model JTT +G.

Protein family	MF	RO	SA	BM
Glucokinases	−77.2	−76.8	−117.4	*−223.6*
Homogentisate 1,2-dioxygenases	−88.9	−62.6	−61.6	*−210.0*
Cytochrome P450	−141.1	−59.2	−106.4	*−249.2*
Pancreatic ribonucleases	−29.5	−13.1	−23.4	−26.2

NOTE.—The values for the latter models were collected from Bordner and Mittelmann (2013). More negative values indicate better models, and the best model is indicated in italics.

Table 3. Difference of AIC, ΔAIC, for the Structurally Constrained Substitution Model MF Relative to the Empirical Substitution Model WAG for Protein Families Filtered at Different Levels of Sequence Identity with Respect to the Protein of the Reference Structure.

Protein family	seq. id. > 0.25	seq. id. > 0.35	seq. id. > 0.45
Triosephosphate isomerases	−121.0 (53)	−57.8 (35)	−21.1 (4)
Rubredoxins	−54.7 (39)	−51.3 (33)	−51.9 (29)
Kinesins	*113.9* (85)	−37.4 (30)	−42.9 (6)
Ferredoxins	−99.5 (25)	−78.8 (24)	−114.8 (19)
DNA ligases	−414.6 (124)	−443.4 (113)	−367.0 (104)
Heat shock proteins	*114.1* (32)	−9.3 (30)	−41.4 (28)
Oxysterol-binding proteins	*118.9* (26)	−24.5 (22)	−60.1 (17)
Retroviral aspartil proteases	*30.1* (18)	−3.2 (3)	NA (2)

NOTE.—Results for data sets where the empirical model better fits the data are shown in italics. In parenthesis, the sample size of such a data set is specified. Note that smaller sample sizes lead to lower absolute ML values and therefore could lead to higher (less negative) ΔAIC scores.

higher than 0.35 presented a better fitting with the MF model than with the empirical model.

Discussion and Conclusions

It is known that the rate at which an amino acid site experiences change is altered by substitutions at neighboring sites due to structural constraints (Liberles et al. 2012; Wilke 2012). Models of evolution that incorporate structural constraints are therefore of increasing importance but, due to their intrinsic complexity, they have not yet been incorporated into the commonly used phylogenetic inference frameworks. This is because the common design of a likelihood function requires site-independent matrices of substitution (Felsenstein 1973, 2004).

Starting from a previous proposal from one of the authors and coworkers (Porto et al. 2005), in this article we have presented a new model for analytically computing site-specific amino acid profiles for proteins of known structure that take into account selection for the folding stability of the experimentally known native state. With respect to our previous work (Porto et al. 2005) based on the Principal Eigenvector (Bastolla et al. 2005) and on the Effective Connectivity (Bastolla et al. 2008) of the contact matrix, the present model implements two main improvements: 1) The algorithm constrains the difference in free energy between the native state and the misfolded state, represented through a simple statistical mechanical model (Minning et al. 2013) and 2) all the parameters are fixed through an ML criterion, with the aim that the model optimally represents observed protein structures.

Although some models of protein evolution consider the misfolded state, this is computationally cumbersome and it is made at the cost of approximations such as considering only maximally compact structures on the cubic lattice (Gutin et al. 1995) or generating misfolded conformations through threading (Bastolla et al. 2003; Goldstein 2011). Applying an analytic, although approximate, treatment of the misfolded state was crucial for its incorporation in the MF model. In addition, we do not know any other model of the substitution process for phylogenetic inference that considers the misfolded state.

Stability against misfolding is thought to be an important requirement in protein evolution. For instance, one of us and coworkers have shown through computational predictions of the stability of orthologous proteins (Bastolla et al. 2004) and through simulations (Mendez et al. 2010) that the interplay between the stability against unfolding and against misfolding is modulated by the mutation process and plays an important role in protein evolution. The results presented here show that considering the stability against misfolding improves the performances of the MF model with respect to a model in which only the native state is considered, because it provides larger likelihood to the observed protein sequences, it avoids that the hydrophobicity is overestimated, and it generates more stable protein sequences.

Besides representing misfolding, our model has another advantage with respect to other models of structural constrained protein evolution such as Rodrigue et al. (2005) and Bordner and Mittelmann (2014). These models approximate the amino acid distributions through pairwise terms, and therefore they cannot be implemented in standard programs of phylogenetic inference, whereas our model with independent sites is much simpler from a computational point of view and it can be combined with standard molecular evolution algorithms.

The method presented here has still a considerable room for improvement, in particular improving two key ingredients of our method that ultimately stem from the mutation model: The exchangeability matrix and the background distribution of amino acids.

The exchangeability matrix very strongly affects the values of the likelihood. We cannot adopt empirical exchangeability matrices such as JTT (Jones et al. 1992) and WAG (Whelan and Goldman 2001), because they represent both mutation and selection, whereas in our model selection is represented by the condition on $\overline{\Delta G}$. Consistently, if we adopt the exchangeability matrix of JTT or WAG together with our MF distributions, we get results that are worse than with the pure empirical models. We addressed this problem by adopting an exchangeability matrix that, together with the MF distributions, produces a flux of amino acids that is

equivalent to the corresponding flux in the empirical model (we impose this condition because the parameters of the empirical models are obtained by estimating fluxes between amino acids). Nevertheless, the performances might improve considerably if we optimize the 190 parameters of the exchangeability matrix for the MF model using a large data set of aligned proteins, as BM did for their structurally constrained model.

An attractive possibility is to derive the exchangeability matrix from an underlying mutation model. We developed a mutation model at the codon level with the double goal to derive an exchangeability matrix $E_{ab}^{\mathrm{MF,mut}}$ devoid of the influence of the selection process that affects empirical exchangeability matrices, and to model a background distribution of amino acids P_a^{mut} with fewer than 19 free parameters. The model that we implemented considered a mutation process at the nucleotide level, the known enhancement of the mutation rate at CpG dinucleotides, and assumed that mutations to stop codons are strongly forbidden by natural selection (this was the only point at which selection entered the model). The parameters of the model were fixed through an ML procedure. Nevertheless, the AIC obtained with the background distributions derived from the mutation model was clearly worse than the one obtained with the frequencies derived from the alignments for all studied families, despite having 4 instead of 19 free parameters. Furthermore, the exchangeability matrix derived from the mutation model had poor performances in terms of likelihood. These results indicate that the mutation model that we applied was not sufficiently accurate with respect to empirical models with more parameters. However, we think that the difficult goal to obtain a better mutation model can be greatly rewarding. Because the requirements that this model has to fulfill to improve the likelihoods are highly demanding, their accomplishment may also yield interesting insight on protein evolution.

Although we only tested its performances for phylogenetic inference, the MF model may have as well applications in the context of protein sequence design, because sequences generated with the model are predicted to correspond to stable proteins, and of protein alignments, given its analogy with Hidden Markov Models.

It is remarkable that, despite the simplicity of the independent sites assumption, the MF model apparently performs better than the method of RO as reported in Bordner and Mittelmann (2014), which uses pairwise distributions. Note that the MF model and the RO model are quite similar under the point of view of parameters, because they both adopt the same empirical exchangeability matrix (JTT) and the same contact interaction energies (Bastolla et al. 2001). Therefore, their differences can be mainly attributed to three points: 1) Including (MF) or not (RO) stability against misfolding; 2) adapted exchangeability matrix (eq. 6) (MF) versus empirical exchangeability matrix (RO); and (3) independent sites (MF) versus pairwise (RO) approximation.

Furthermore, in three of the four cases the MF model performs better than the independent sites version of the method of BM that is based on the solvent accessibility of each site (SA) and in principle is similar to our method, despite the SA method optimizes a large number of parameters from databases of protein families. Then, in one over four cases, MF model performs better than the new pairwise model by Bordner and Mittelmann (2014) that is based on factor graphs and is computationally much more complex than the MF model and optimizes for phylogenetic inference parameters that are equivalent to a contact interaction matrix and an exchangeability matrix.

It is interesting that in all three cases in which the BM model outperforms the MF model, the sequence identity is larger than 90%, while the MF is the best one when the average sequence identity drops to the (still high) value of 74%. Based on few comparisons, we do not know whether this behavior is general; however, it suggests that the advantage of using pairwise distributions instead of independent sites does not increase for highly divergent sequences, as one might have expected, because the independent sites approximation is only accurate at small evolutionary distances.

Methods

Background Distribution

The first ingredient of our MF model is the site-independent background amino acids distribution P_a^{mut} that we attribute to the underlying mutation process. We obtained the best results when these frequencies are derived from the frequencies observed in the PDB sequence or in the multiple alignment. In this case, the background distribution has 19 free parameters. All results presented in the Results section were obtained with this choice. We also tried to reduce the number of free parameters defining a mutation process at the codon level. This attempt gave poor results, but it may be an important direction of future improvement. Another possible direction for improvement would be to weight the sequences in the multiple alignment in order to reduce the influence on the background distribution of unbalanced phylogenetic sampling.

Folding Free Energy

We adopt a model of protein folding stability based on contact interactions. We consider three thermodynamic states: The native state, which is assumed to consists of a folded structure with its attraction basin, a state consisting of misfolded compact conformation, and the unfolded state. The vibrational entropy (entropy of the protein confined to its local energy minimum, such that it can be computed through normal mode analysis of the native state or a particular misfolded state) of the folded native state is assumed to be compensated by the vibrational entropy of each misfolded state (Karplus et al. 1987), therefore it is not estimated. We estimate the native free energy as

$$G_{\mathrm{nat}}(C^{\mathrm{nat}}, A) \approx \sum_{i<j} C_{ij}^{\mathrm{nat}} U(A_i, A_j)\,, \qquad (9)$$

where C_{ij}^{nat} is the contact matrix of the native structure represented in the PDB ($C_{ij}^{\mathrm{nat}} = 1$ if residues i and j are closer than 4.5 Å, 0 otherwise), A_i is the amino acid at position i, and

$U(a, b)$ is the 20×20 contact interaction matrix of Bastolla et al. (2001). The free energy of the unfolded state is estimated as $G_U \approx -TLS_U$, where T is the temperature in units in which $k_B = 1$, L is the chain length, and S_U is the conformational entropy per residue of an unfolded chain. The misfolded state consists of the ensemble of compact but wrongly folded conformations, which we model as an ensemble of contact matrices of length L and number of contacts in the range expected for compact protein structures, whose statistical properties are obtained analyzing the compact submatrices of L residues in the PDB, a technique designated as threading in the bioinformatics jargon. Its statistical mechanics is often described by the random energy model (Derrida 1981) that models the energy as a Gaussian random variable (Garel and Orland 1988; Shakhnovich and Gutin 1989; Bryngelson et al. 1995), so that the free energy is determined by the first and second moment of the energy. A more accurate computation also includes the third moment of the energy (Minning et al. 2013). We implemented this correction, but we found that it slightly worsens the likelihood, perhaps due to the approximations that we have to adopt for making the iterative computation feasible, and we do not consider it in our default algorithm, which is based on the following model of the free energy of the misfolded state:

$$G_{\mathrm{misf}} \approx \sum_{i<j} \langle C_{ij} \rangle U(A_i, A_j) - \frac{1}{2T} \sum_{i<j,k<l} \left(\langle C_{ij} C_{kl} \rangle - \langle C_{ij} \rangle \langle C_{kl} \rangle \right) U(A_i, A_j) U(A_k, A_l) - LS_C T, \tag{10}$$

where LS_C is the logarithm of the number of compact contact matrices, $\langle . \rangle$ represents the average over the set of compact contact matrices of L residues, and we assume for simplicity that the conformational entropy $S(C_{ij})$ is approximately the same for all compact structures and it can be ignored for computing free energy differences.

Our computational problem is to compute the sequence average of equation (10) in a way that is fast enough for allowing several iterations of the MF algorithm. For this reason, we simplify the computation of the misfolding free energy as detailed in the next section.

Solution of the Mean-Field Model

By equating the derivatives of equation (3) to zero, we obtain the following implicit solution of the MF equation:

$$P_a^{\mathrm{MF},\Lambda,i} = z_i P_a^{\mathrm{mut}} \exp\left(-\Lambda \frac{\partial \overline{\Delta G}}{\partial P_a^{\mathrm{MF},\Lambda,i}} \right), \tag{11}$$

$$\overline{\Delta G} = \sum_{a_1 \cdots a_L} P_{a_1}^{\mathrm{MF},\Lambda,1} \cdots P_{a_L}^{\mathrm{MF},\Lambda,L} (G_{\mathrm{nat}}(a_1 \cdots a_L) - G_{\mathrm{misf}}(a_1 \cdots a_L)), \tag{12}$$

where a denotes one of the 20 amino acids, i is a protein site, z_i is determined through the normalization condition $\sum_a P_a^{\mathrm{MF},\Lambda,i} = 1$, and $\overline{\Delta G}$ is the MF average of the folding free energy. Starting from an initial guess or from the distribution previously obtained for a close value of Λ, these equations are iterated until convergence. However, because convergence is not guaranteed, after a large number of iterations our algorithm chooses the distribution closest to convergence. We observed that this criterion yields the largest final likelihood. The above equations are explicitated in the Appendix, where we describe all necessary computations.

ML Optimization of the Lagrange Multiplier

In order to numerically determine the value of Λ that maximizes the likelihood, we compute the MF distribution for values of Λ, starting from $\Lambda = 0.1$ and incrementing it by 0.1 at each step. The solution $P_a^{\mathrm{MF},\Lambda,i}$ relative to the previous value of Λ is used as the starting point of the iterative algorithm. After this coarse exploration, the value of Λ that maximizes the likelihood is obtained through iterative quadratic interpolations.

Methods for Phylogenetic Inference

In order to apply the MF model for phylogenetic inference, we input to our program Prot_Evol the representative protein structure and all the sequences of the protein family and we obtain as output the site-specific amino acid distributions and the global exchangeability matrix. We then align the sequences with the program MAFFT (Katoh and Standley 2013). This was done even for families that were aligned in the Pfam database (Punta et al. 2012), because we observed that realigning them improved the quality of the alignment and the values of the likelihood for all models. We discard columns of the alignment for which the representative protein or more than 50% of the proteins have a gap, and we compute a phylogenetic tree applying the Neighbor Joining algorithm of Saitou and Nei (1987).

The alignment, the tree, and the substitution models (either empirical substitution models or the model generated with the MF distributions) are then input to the program PAML (Yang 2007) for computing the likelihood of the data given in the model. For the site-specific MF models, we proceed in two steps. In the first step, we optimize the branch lengths for all sites using the complete alignment and the site-averaged amino acid frequencies, $\sum_i P_a^{\mathrm{MF},i}/L$. In the second step, we compute the likelihood for each site using the corresponding column of the alignment, the site-specific frequencies, and the branch lengths optimized in the previous step. We use the same exchangeability matrix in both steps. The computation of branch lengths is required to modify the code of PAML, because this program internally normalizes the rate matrix in such a way that the average rate is always one. In this way, the time unit of the rate matrix is lost and the branch lengths are output in arbitrary units, which would prevent using them in the second step. To avoid this problem, we eliminated the internal normalization of the rates.

Acknowledgments

U.B. gratefully thanks Markus Porto for participating in many prior stages of this work. We would like to thank David Liberles for helpful comments and discussions. This work

was supported by the Spanish Ministery of Economy through the grant BFU-40020 to U.B. M.A. was supported by the Spanish Government through the Juan de la Cierva fellowship JCI-2011-10452. Research at the CBMSO is facilitated by the Fundación Ramón Areces. We thank three anonymous reviewers for insightful comments.

Appendix: Numerical Solution of the Mean-Field Equations

We start by approximating the misfolding free energy through the average contact energy of the misfolding ensemble, so that it holds $\Delta G \approx \sum_{ij}(C_{ij}^{\rm nat} - \langle C_{ij}\rangle)U(A_i, A_j) + TS_C$. This approximation was indicated as $P^{{\rm MF1},\Lambda,i}$ in the main text. In this case, equation (11) has the simple form

$$P_a^{{\rm MF1},\Lambda,i} \propto P_a^{\rm mut}\exp\Big(-\Lambda\Big(C_{ij}^{\rm nat} - \langle C_{ij}\rangle\Big)u_{ja}\Big), \qquad (13)$$

where $u_{ja} = \sum_b P_b^{{\rm MF1},\Lambda,j}U(a,b)$ is the interaction energy of amino acid a interacting with the ensemble of amino acids present at site j, in the spirit of the MF approximation. We can further simplify it by adopting the hydrophobic approximation $U(a,b) \approx \epsilon_0 + \epsilon h_a h_b$, where the so-called hydrophobicity vector h_a is determined as the main eigenvector of the contact interaction matrix $U(a,b)$ [38]. In this case, we obtain

$$P_a^{{\rm MF1},\Lambda,i} \propto P_a^{\rm mut}\exp\Big(-\Lambda h_a\Big(C_{ij}^{\rm nat} - \langle C_{ij}\rangle\Big)H_j\Big), \qquad (14)$$

where the field $H_j = \sum_b P_b^{{\rm MF1},\Lambda,j}h_b$ can be interpreted as the MF hydrophobicity of site j. The self-consistent equation (14) can be solved iteratively and they rapidly converge. They can be used as a starting point for the more complicated MF models that include other terms of the misfolding free energy. In the following, we simplify the notation by omitting the superscript Λ, with the understanding that this parameter is fixed at the value determined by the ML condition (eq. 4).

If we set $\langle C_{ij}\rangle = 0$ in the above equations, considering only the native free energy, we obtain the zeroth order MF model $P_a^{{\rm MF0},i}$ that is qualitatively similar to an earlier proposal by one of us and coworkers (Porto et al. 2005). The models $P_a^{{\rm MF2},i}$ and $P_a^{{\rm MF3},i}$ are obtained by adding the second and the third moment of the misfolding energy, respectively. We found the best results with $P_a^{{\rm MF2},i}$, which we will denote as $P_a^{{\rm MF},i}$, omitting the superscript that specifies the order of the misfolding free energy.

When including the second moment of the energy, we have to consider the correlations between pairs of contacts, whose number grows as the fourth power of the number of sites L. There is not enough data to accurately compute such correlations, and storing this information in memory would cause computational problems. Therefore, we reduce the size of the data that have to be estimated and stored adopting the so-called homogeneous approximation (Minning et al. 2013). This approximation assumes that the probability of a contact between two sites only depends on their difference in the sequence but not on their absolute position, $\langle C_{ij}\rangle \approx f(|i-j|)$, so that the number of data increases only linearly with L. For estimating the contact correlations $D_{ijkl} = \langle C_{ij}C_{kl}\rangle - \langle C_{ij}\rangle\langle C_{kl}\rangle$, we have to distinguish three cases: 1) $ij = kl$, that is, only two of the sites are different; we indicate the corresponding contact correlation as $D_{ijij} \approx (C221)_{|i-j|}$, where the numbers indicate that this is the contact correlation of order 2 with 2 different sites and 1 different contact; 2) $i = k$, $j \neq l$, that is, three of the sites are different; we approximate $D_{ijil} \approx (C232)_i$ neglecting the dependence on sites j and l; 3) all four sites are different, $D_{ijkl} \approx (C242)$, and in this case we neglect the dependence on all four indices. These coefficients are estimated as

$$(C221)_{ij} = \frac{1}{2}\left(\langle C_{ij}\rangle - \langle C_{ij}\rangle^2\right) \qquad (15)$$

$$(C232)_i = \frac{1}{4}\frac{\langle m_i^2\rangle - \langle m_i\rangle - \left(\langle m_i\rangle^2 - \sum_j\langle C_{ij}\rangle^2\right)}{\langle m_i\rangle^2 - \sum_j\langle C_{ij}\rangle^2} \qquad (16)$$

$$(C242) = \frac{\langle N_c^2\rangle - \sum_i\langle m_i^2\rangle - \left(\langle N_c\rangle^2 - \sum_i\langle m_i\rangle^2\right)}{\langle N_c\rangle^2 - \sum_i\langle m_i\rangle^2}, \qquad (17)$$

where m_i is the number of contacts of site i, $\langle m_i\rangle$ is its average over the misfolding ensemble, and $N_c = \frac{1}{2}\sum_i m_i$ is the total number of contacts. With this notation, we compute the second moment of the energy of the misfolded ensemble as

$$\begin{aligned}\langle E^2\rangle - \langle E\rangle^2 &= \sum_{i<j,k<l} D_{ijkl}U_{ij}U_{kl}\\ &\approx \sum_{ij}(C221)_{ij}(U_{ij})^2 + \sum_i\sum_{j\neq l}(C232)_i\langle C_{ij}\rangle\langle C_{il}\rangle U_{ij}U_{il}\\ &+ \sum_{ijkl\ \text{diff}}(C242)U_{ij}U_{kl},\end{aligned} \qquad (18)$$

and the MF distribution can be computed as

$$P_a^{{\rm MF2},\Lambda,i} \propto P_a^{\rm mut}\exp(-\Lambda H_{ia}) \qquad (19)$$

$$\begin{aligned}H_{ia} &= \Big(C_{ij}^{\rm nat} - \langle C_{ij}\rangle\Big)u1_{ja}\\ &+ \frac{1}{2T}\left[\sum_j(C221)_{ij}u2_{ja} + (C232)_i\left((g1_{ia})^2 - g2_{ia}\right)\right]\\ &+ \frac{1}{T}\sum_k\left[(C232)_k\langle C_{ik}\rangle(ug1_{ka} - uu1_{ka}) + (C242)\overline{\langle E\rangle}g1_{ka}\right]\end{aligned} \qquad (20)$$

with $u1_{ja} = \sum_b P_b^{{\rm MF2},\Lambda,j}U(a,b)$, $u2_{ja} = \sum_b P_b^{{\rm MF2},\Lambda,j}U(a,b)^2$,

$g1_{ia} = \sum_j \langle C_{ij} \rangle u1_{ja}$, $g2_{ia} = \sum_j \left(\langle C_{ij} \rangle u1_{ja} \right)^2$, $ug1_{ka} = \sum_b P_b^{MF2,\Lambda,j} U(a,b) g1_{kb}$, $uu1_{ka} = \sum_b P_b^{MF2,\Lambda,j} U(a,b) u1_{kb}$ and $\langle E \rangle = \sum_{i<j} \langle C_{ij} \rangle \sum_b P_a^{MF2,\Lambda,i} \sum_b P_b^{MF2,\Lambda,j} U(a,b)$.

Finally, we have to take into account that the analytic expression for computing the free energy of the misfolded state, equation (10), is only valid if the temperature is higher than the freezing temperature of the system, $T_f = \sqrt{\langle (E - \langle E \rangle)^2 \rangle / 2S_C}$. For the MF model, we determine the freezing temperature using the average of the second moment of the energy over the MF distribution, $\overline{\langle (E - \langle E \rangle)^2 \rangle}$. If the temperature T is smaller than the freezing temperature, we have to use T_f instead of T in equation (20).

References

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automatic Control.* 19:716–723.

Arenas M, Dos Santos HG, Posada D, Bastolla U. 2013. Protein evolution along phylogenetic histories under structurally constrained substitution models. *Bioinformatics* 29:3020–3028.

Babajide A, Hofacker IL, Sippl MJ, Stadler PF. 1997. Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold Des.* 2:261–269.

Bastolla U, Farwer J, Knapp EW, Vendruscolo M. 2001. How to guarantee optimal stability for most representative structures in the Protein Data Bank. *Proteins* 44:79–96.

Bastolla U, Moya A, Viguera E, van Ham RC. 2004. Genomic determinants of protein folding thermodynamics in prokaryotic organisms. *J Mol Biol.* 343:1451–1466.

Bastolla U, Ortiz AR, Porto M, Teichert F. 2008. Effective connectivity profile: a structural representation that evidences the relationship between protein structures and sequences. *Proteins* 73:872–888.

Bastolla U, Porto M, Roman HE, Vendruscolo M. 2003. Statistical properties of neutral evolution. *J Mol Evol.* 57(Suppl 1), S103–S119.

Bastolla U, Porto M, Roman HE, Vendruscolo M. 2005. Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins* 58:22–30.

Bastolla U, Porto M, Roman HE, Vendruscolo M. 2006. A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. *BMC Evol Biol.* 6:43.

Bastolla U, Roman HE, Vendruscolo M. 1999. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J Theor Biol.* 200:49–64.

Berg J, Willmann S, Lässig M. 2004. Adaptive evolution of transcription factor binding sites. *BMC Evol Biol.* 4:42.

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. 2005. Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci U S A.* 102:606–611.

Bordner AJ, Mittelmann HD. 2014. A new formulation of protein evolutionary models that account for structural constraints. *Mol Biol Evol.* 31:736–749.

Bornberg-Bauer E, Chan HS. 1999. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci U S A.* 96:10689–10694.

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* 21:167–195.

Bussemaker HJ, Thirumalai D, Bhattacharjee JK. 1997. Thermodynamic stability of folded proteins against mutations. *Phys Rev Lett.* 79:3530–3533.

Chan HS, Zhang Z, Wallin S, Liu Z. 2011. Cooperativity, local-nonlocal coupling, and nonnative interactions: principles of protein folding from coarse-grained models. *Annu Rev Phys Chem.* 62:301–326.

DePristo MA, Weinreich DM, Hartl DL. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet.* 6:678–687.

Derrida B. 1981. Random energy model: an exactly solvable model of disordered systems. *Phys Rev B.* 24:2613–2626.

Dokholyan NV, Shakhnovich EI. 2001. Understanding hierarchical protein evolution from first principles. *J Mol Biol.* 312:289–307.

Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst Zool.* 22:240–249.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Fornasari MS, Parisi G, Echave J. 2002. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Mol Biol Evol.* 19:352–356.

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol.* 26:2387–2395.

Garel T, Orland H. 1988. Mean-field model for protein folding. *Europhys Lett.* 6:307–310.

Goldstein RA. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* 79:1396–1407.

Govindarajan S, Goldstein RA. 1997. Evolution of model proteins on a foldability landscape. *Proteins* 29:461–466.

Grahnen JA, Nandakumar P, Kubelka J, Liberles DA. 2011. Biophysical and structural considerations for protein sequence evolution. *BMC Evol Biol.* 11:361.

Gutin AM, Abkevich VI, Shakhnovich EI. 1995. Evolution-like selection of fast-folding model proteins. *Proc Natl Acad Sci U S A.* 92:1282–1286.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol.* 15:910–917.

Huang TT, del Valle Marcos ML, Hwang JK, Echave J. 2014. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. *BMC Evol Biol.* 14:78.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8:275–282.

Karplus M, Ichiye T, Pettitt BM. 1987. Configurational entropy of native proteins. *Biophys J.* 52:1083–1085.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780.

Koshi JM, Goldstein RA. 1998. Models of natural mutations including site heterogeneity. *Proteins* 32:289–295.

Koshi JM, Goldstein RA. 2001. Analyzing site heterogeneity during protein evolution. *Pac Symp Biocomput.* 191–202.

Koshi JM, Mindell DP, Goldstein RA. 1999. Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. *Mol Biol Evol.* 16:173–179.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21:1095–1109.

Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, de Koning AP, Dokholyan NV, Echave J, et al. 2012. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.* 21:769–785.

Mendez R, Fritsche M, Porto M, Bastolla U. 2010. Mutation bias favors protein folding stability in the evolution of small populations. *PLoS Comput Biol.* 6:e1000767.

Minning J, Porto M, Bastolla U. 2013. Detecting selection for negative design in proteins through an improved model of the misfolded state. *Proteins* 81:1102–1112.

Mirny LA, Abkevich VI, Shakhnovich EI. 1998. How evolution makes proteins fold quickly. *Proc Natl Acad Sci U S A.* 95:4976–4981.

Parisi G, Echave J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. *Mol Biol Evol.* 18:750–756.
Plotkin SS, Onuchic JN. 2002. Understanding protein folding with energy landscape theory. Part II: quantitative aspects. *Q Rev Biophys.* 35: 205–286.
Porto M, Roman HE, Vendruscolo M, Bastolla U. 2005. Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. *Mol Biol Evol.* 22: 630–638.
Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. 2012. The Pfam protein families database. *Nucleic Acids Res.* 40:D290–D301.
Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol.* 20:1692–1704.
Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* 347:207–217.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.
Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci U S A.* 102:9541–9546.
Shakhnovich E. 2006. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem Rev.* 106: 1559–1588.
Shakhnovich EI, Gutin AM. 1989. Formation of unique structure in polypeptide chains. *Biophys Chem.* 34:187–199.
Taverna DM, Goldstein RA. 2002. Why are proteins marginally stable? *Proteins* 46:105–109.
Tiana G, Broglia RA, Roman HE, Vigezzi E, Shakhnovich EI. 1998. Folding and misfolding of designed proteinlike chains with mutations. *J Chem Phys.* 108:757–761.
Uversky VN, Dunker AK. 2010. Understanding protein non-folding. *Biochim Biophys Acta.* 1804:1231–1264.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol.* 18:691–699.
Whelan S, Liò P, Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* 17: 262–272.
Wilke CO. 2012. Bringing molecules back into molecular evolution. *PLoS Comput Biol.* 8:e1002572.
Yang Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol.* 10: 1396–1401.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
Yang Z, Nielsen R, Masami H. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol.* 15:1600–1611.